基于递进式卷积网络的农业命名实体识别方法

计　洁[1,2]，金　洲[1]，王儒敬[1,2*]，刘海燕[1,2]，李志远[1,2]

（1. 中国科学院合肥物质科学研究院智能机械研究所，安徽合肥 230031；2. 中国科学技术大学，安徽合肥 230026）

摘　要：目前基于预训练语言模型（Pre-trained Language Model，PLM）的命名实体识别的研究在面对农业领域存在的实体命名方式繁杂、实体边界模糊等问题时，仅使用PLM最后一层表示输出，且均从外部引入知识或操作对实体表示进行增强，忽视内部各层本身蕴含语言不同层次的丰富信息。为解决上述问题，提出一种基于递进式卷积网络的命名实体识别方法。该方法首先存储自然句子，通过PLM后得到的每层输出表示；其次以递进式卷积作为全层信息的特征提取手段，对储存的模型中间层输出表示依次卷积。模型将注重全层信息，包括被忽略的浅层输出，而有研究表明靠近输入的模型层输出的句子嵌入包含更多的诸如短语、词组等粗粒度信息，对于边界模糊的农业命名实体识别，更关键的词组界定信息或许就隐含在这些被忽略的浅层嵌入中，可为农业领域存在的命名实体识别问题提供帮助。无需外部信息的引入，充分利用已使用的计算力得到的结果就能增强句子的表示嵌入；最终通过条件随机场（Conditional Random Field，CRF）模型生成全局最优序列。在构建的包含农作物品种、病害、虫害和农药4类农业实体的农业数据集上，所提方法的综合性指标 F_1 值相较于基于Transformer的双向编码表征模型（Bidirectional Encoder Representation from Transformers，BERT）提升3.61%，在公开数据集上也有较好表现，其中在数据集MSRA上 F_1 值提升至94.96%，说明基于递进式的卷积网络能够增强模型对自然语言的表示能力，在命名实体识别任务上具有优势。

关键词：农业命名实体识别；预训练语言模型；卷积网络；表示聚合；深度学习

中图分类号：TP391.1　　**文献标志码：**A　　**文章编号：**SA202303001

引用格式：计洁, 金洲, 王儒敬, 刘海燕, 李志远. 基于递进式卷积网络的农业命名实体识别方法[J]. 智慧农业(中英文), 2023, 5(1): 122-131.

JI Jie, JIN Zhou, WANG Rujing, LIU Haiyan, LI Zhiyuan. Progressive convolutional net based method for agricultural named entity recognition[J]. Smart Agriculture, 2023, 5(1): 122-131.

1　引　言

随着农业信息化水平的提高，网络中的农业数据也以指数规模增长，利用这些数据可以对知识问答、知识图谱构建等农业知识服务进行研究。在非结构化文本占大多数的农业数据中识别出有意义的名词或短语并加以归类，如农作物品种、病害、虫害和农药名称等，即农业命名实体识别，是上述知识服务的关键性环节，也是后续获取高质量的语义知识、进行农业信息抽取与语义检索的支撑。提升命名实体识别（Named Entity Recognition，NER）的准确率可以有效地为农业领域进行知识服务，让农业决策者、生产经营者及科研人员及时准确地获取到覆盖完整、高度相关的信息，提升农业整体发展水平。

预训练的出现给自然语言处理带来新的生机[1,2]，现较有效的命名实体识别方法均基于预训练语言模型实现。特别地，基于Transformer的双向编码表征（Bidirectional Encoder Representation from Transformers，BERT）模型[3]近些年被广泛

收稿日期：2023-03-03

基金项目：国家重点研发计划项目（2019YFE0125700）

作者简介：计　洁，硕士研究生，研究方向为自然语言处理、知识图谱。E-mail：jijiejie@mail.ustc.edu.cn

*通信作者：王儒敬，博士，研究员，研究方向为智能决策与知识工程、专家系统。E-mail：rjwang@iim.ac.cn

应用到开放领域和垂直领域的命名实体识别中。BERT模型是由多层Transformer[4]堆叠而成的深度模型，具有学习上下文信息的编码能力。这些语言模型利用注意力机制学习上下文信息，将自然语言转换为语义信息丰富的句子嵌入。利用这种特性，杨飘和董文永[5]提出使用BERT来优化传统的双向门控循环单元+条件随机场（Bidirectional Gating Recurrent Unit + Conditional Random Field ， BiGRU + CRF）方法，实现中文命名实体识别。Gan等[6]将BERT结合双向长短期记忆网络（Bi-directional Long Short-Term Memory， BiLSTM）和条件随机场（Conditional Random Field， CRF）， Gao等[7]则将此种框架结合成中文命名实体识别算法模型，应用于CCKS2020电子病历数据集的处理。Chang等[8]同时使用BiLSTM和迭代膨胀卷积神经网络（Iterated Dilated Convolutional Neural Networks， IDCNN）对句子嵌入做特征提取，再将两种提取的特征行融合，获得表示能力更强的句子嵌入。Li等[9]在BERT输入引入外部实体知识，将传统的lattice结构展开为平面结构。另一方面，一些工作关注于如何从模型输入端增强句子中实体之间的关联性。琚天根等[10]提出了基于关联记忆网络的中文命名实体识别方法。Wang等[11]提出将搜索引擎查询获得的与输入句子匹配度较高的文本和输入句子一起经过预训练语言模型获得嵌入，实现实体表示增强的效果。Nie等[12]提出利用键值记忆网络进行信息编码。

在农业领域，存在实体名称组成繁杂、长度较长[13]、词语边界模糊、罕见词识别率低等问题[14]。但上述的命名识别方法都侧重于引入外部知识或者操作来增强句子特征，以增加下游模型复杂度来换取精度的提升，对增强信息进行舍近求远的操作；同时在对齐外部引入数据时使得模型通用性较差。而相关研究表明[2, 15, 16]，BERT每一层能够学习到不同维度的语言学信息：靠近输入的Transformer层输出的句子嵌入包含更多的诸如短语、词组等信息；更深的编码层会偏向句子的语义信息。对于农业命名实体识别而言，短语、句法等更粗粒度的信息可能会更有用[17]。若在使用时只提取预训练语言模型的最后一层表示，可能会错过更利于命名实体识别的浅层信息。例如，关于农业知识图谱问答中，用户问句往往较为简短，拥有典型句子成分。如用户问句“小麦白粉病会发生于哪些作物上？”中，“小麦”和“白粉病”都可以视为独立的实体，而在农业领域中理解此问句时，更希望得到“小麦白粉病”这一完整的命名实体。因此，对于农业知识图谱问答中的命名实体识别任务来说，利用粗粒度的语言信息会更好地把握具有较长名称的农业实体名称，发挥重要作用。如果仅提取预训练语言模型的最后一层表示，可能会错过这些更利于命名实体识别的浅层信息。特别是对于边界模糊的农业命名实体识别任务，更关键的词组界定信息或许就隐含在这些被忽略的浅层嵌入中。因此，只使用最后一层输出当作句子表示是存在一定缺陷的。

在机器翻译领域，Zhang等[18]提出BERT-JAM模型对BERT多层进行融合。BERT-JAM首先给模型各层赋予可训练的权重，然后利用门控单元对各层表示进行选择。另外，Su和Cheng[19]提出一种基于Squeeze和Excitation[20]的BERT多层表示加权的模型SesameBERT，在句子分类和自然语言推断等多个数据集上有提升效果。但在命名实体识别领域针对预训练语言模型的多层表示融合研究较少。

为了解决农业领域中存在的实体识别问题，充分挖掘预训练语言模型内部不同维度的语言学信息，本研究提出一种结合预训练语言模型和递进式卷积网络的命名实体识别方法，利用递进式卷积网络对模型各层表示进行聚合，无需添加外部信息，仅使用卷积去深度利用模型多层表示蕴含的丰富信息。由于采用的是递进式的卷积，它不仅保存模型深层输出所拥有的细粒度语义信息，同时注重浅层的粗粒度信息，在更好地把握实体名称的长度界定信息下，将繁杂的实体作为整体，去融合和探究实体处于句子中的上下文信息与位置，会比仅仅通过细粒度地分析语义更适合如农业等专业领域的命名实体识别。Jiang等[21]阐述由于大型深度模型很大程度上依赖于全局自我注意块而会产生冗余的信息，同时从侧面证明卷积这一提取特征的有效工具可以与自我注意块忽略的信息进行互补性探究。将预训练语言模型的中间层输出依次进行卷积，可以提取原先可能被忽视的浅层特征信息，在无需外部

信息的引入下，充分利用已使用的计算力得到的结果来增强自然语言的表示嵌入。实验结果表明，本研究所提方法对多个预训练语言模型均有效，在多个命名实体识别数据集上的准确率均得到提升。

2 基于递进式网络的表示融合模型构建与训练

本研究提出一种结合预训练语言模型和递进式卷积网络的命名实体识别方法，使用卷积操作来挖掘预训练语言模型所有编码层的信息，对所有层输出集合进行递进形式的卷积，即将相邻每两层卷积结果继续与集合中下一层输出编码进行卷积融合，直至模型最后一层。将最终融合编码结果作为输入句子的表示再进行后续命名实体识别解码，方法模型如图1所示。

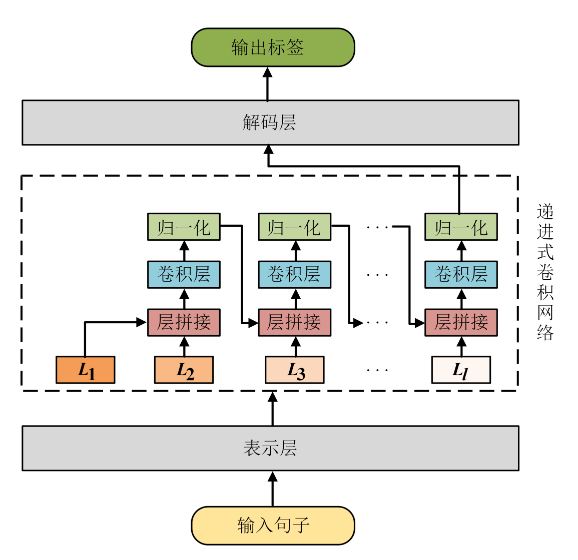


图1 基于递进式网络的表示融合模型整体框架图

Fig. 1 Overall framework of the representation fusion model based on progressive networks

模型结构主要包括表示层、递进式卷积网络和解码层。其中，表示层利用预训练语言模型对输入句子进行编码，将自然语言转换为学习了上下文信息的嵌入表示集合。递进式卷积网络利用卷积层对获得的表示集合进行聚合，卷积模块从第一层开始对相邻两层卷积融合，融合结果作为新的聚合表示继续与下一未卷积层进行卷积融合，故称之为递进式融合。卷积直至表示集合的最后元素，即模型得到的最后一层表示被融合。该操作提取不同维度的语言学信息，获得增强的句子嵌入。增强的嵌入包含模型全层的信息维度，在没有增加外部信息的情况下，使得句子表示包含更丰富的信息。解码层负责对表示解码，利用CRF [22] 学习句子嵌入与对应标签之间的关系，并利用维特比算法输出预测的标签序列。

2.1 表示层

表示层用来将自然语言序列编码为向量表示。在表示层使用预训练语言模型编码输入句子，可以获得学习到上下文信息的嵌入表示集合。以BERT为例，表示层结构如图2所示。其中，Trm表示Transformer [11] 的Encoder部分。

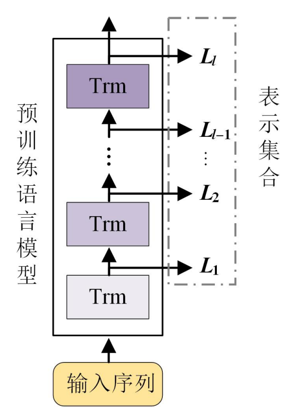


图2 预训练语言模型表示层

Fig. 2 Pre-trained language model presentation layer

给定预训练语言模型PLM，对输入序列进行编码。预训练语言模型为多层Transformer结构，因此，可以得到句子的表示集合 $\boldsymbol{LS}$，如公式（1）所示。

$$\boldsymbol{LS} = \mathrm{PLM}(\boldsymbol{S}) = \{\boldsymbol{L}_1, \boldsymbol{L}_2, ..., \boldsymbol{L}_l\} \tag{1}$$

其中，$\boldsymbol{L}_i \in \boldsymbol{R}^{n\times h}$，$i\in\{1, 2, \cdots, l\}$ 表示预训练语言模型第 i 层编码的句子表示；l 是预训练语言模型深度；$\boldsymbol{R}$ 为实数矩阵集；n、h 分别表示句子长度和预训练语言模型隐层维度。

预训练语言模型不同层编码的表示侧重点各不相同，所以表示层获得的表示集合包含诸如短语、词法、词序、句子语义等多个维度的语言学信息。因此，表示集合 $\boldsymbol{LS}$ 能够更充分地表示输入句子。

2.2 递进式卷积网络构建

区别于现有方法，本研究设计了递进式卷积网络。递进式卷积网络深度为 l–1，每一层结构相同，

如图3所示。

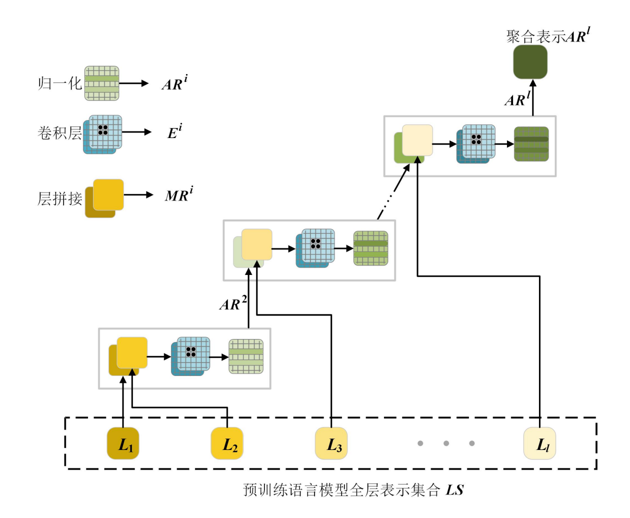


图3 递进式卷积网络结构图

Fig. 3 Network structure of progressive convolutional

网络每层由三部分组成：层拼接、卷积层和归一化。其中，层拼接是将前一层的输出的融合表示$\boldsymbol{AR}^{i-1}\in \boldsymbol{R}^{n\times h}$与当前层的句子嵌入$\boldsymbol{L}_i\in \boldsymbol{R}^{n\times h}$拼接起来得到多维度混合表示$\boldsymbol{MR}^i\in \boldsymbol{R}^{2\times n\times h}$，$i\in\{1，2，\cdots，l\}$。拼接后有利于下一步对两者进行卷积操作，如公式（2）所示。

$$\boldsymbol{MR}^i = concat(\boldsymbol{AR}^{i-1}, \boldsymbol{L}_i) = [\boldsymbol{AR}^{i-1};\ \boldsymbol{L}_i] \quad (2)$$

卷积层的目的是实现前一层输出的融合表示$\boldsymbol{AR}_{i-1}$和当前层句子嵌入$\boldsymbol{L}_i$的融合。即对于递进式卷积网络第c层的卷积层，其输入为$\boldsymbol{MR}^i$，卷积核为$\boldsymbol{k}^c\in\boldsymbol{R}^{2\times w\times b\times 1}$，$w$和$b$分别为卷积核的长和宽，输出为$\boldsymbol{E}^c_{i,\ j}\in\boldsymbol{R}^{w\times b}$。$\boldsymbol{E}^c$的计算过程如公式（3）所示。

$$\begin{cases} \boldsymbol{E}^c = \boldsymbol{MR}^c \otimes \boldsymbol{k}^c \\ \boldsymbol{E}^c_{i,j} = \sum\limits_{x=i-\frac{w}{2}}^{i+\frac{w}{2}} \sum\limits_{y=j-\frac{b}{2}}^{j+\frac{b}{2}} \left(AR^{c-1}_{x,y} \cdot \boldsymbol{k}^{c,0}_{x-i+\frac{w}{2},y-j+\frac{b}{2}} + LR^c_{x,y} \cdot \boldsymbol{k}^{c,1}_{x-i+\frac{w}{2},y-j+\frac{b}{2}} \right) \end{cases} \quad (3)$$

其中，$\boldsymbol{E}^c\in\boldsymbol{R}^{w\times b}$为当前卷积层输出；$LR_{x,\ y}$为当前层的句子表示$x$行$y$列的元素；$E^c_{i,\ j}$为$\boldsymbol{E}^c$矩阵$i$行$j$列的元素；$AR_{x,\ y}$为$\boldsymbol{AR}$矩阵$x$行$y$列的元素。

通过层连接和卷积层，可以保证融合前后句子嵌入的尺寸大小不变。同时相较于融合之前，融合之后的句子嵌入提取了当前层的特征。对于序列中的某一位置而言，卷积操作可以使得其学习到上下文表示的特点，并且其学习到的上下文范围受到卷积核尺寸的影响。对于命名实体识别而言，命名实体是具有一定跨度的字符序列，对于命名实体中的某一字符，学习实体的其他字符的表示有助于实体识别任务。

归一化层的目的是保证当前层卷积后的句子嵌入能够与卷积之前量级上保持一致，有利于下一层的融合，归一化层也给表示带来非线性变化，提高网络对模型的表达能力。不同于计算机视觉领域用批量归一化（Batch Normalization）对批量里的不同样本的同一特征做归一操作，本研究方法利用层归一化（Layer Normalization），针对单个样本的不同特征做归一操作，不会破坏同一句子上下文中不同词义向量的可比性，更适合自然语言处理任务，有助于模型训练，避免过拟合。如图3所示，给定$\boldsymbol{E}^c$，其归一化值$\boldsymbol{AR}^c$由公式（4）～（6）计算得到：

$$u_i = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{E}^c_{j,i} \quad (4)$$

$$\sigma_i = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{E}^c_{j,i} - u_i)^2} \quad (5)$$

$$AR^c_{j,i} = \frac{g_i}{\sigma_i}(\boldsymbol{E}^c_{j,i} - u_i) + b \quad (6)$$

其中，u_i为$\boldsymbol{E}^c$第i列的平均数；n为句子长度；σ_i为第i列的方差；g_i和b是可训练参数。可以看出，该归一化方法在句子嵌入的同一隐层维度上进行归一化。

2.3 解码层

解码层负责将2.2节得到的聚合分布表示$\boldsymbol{AR}^c$解码为对应的标签序列。为了获得最优标签序列，使用CRF [22] 对其进行解码。对于输入的句子序列$\boldsymbol{S}$ $=\{s_1，s_2，\cdots，s_n\}$，其对应的标注序列为$y=\{y_1，y_2，\cdots，y_n\}$，$y_i\in$Y，Y为命名实体的类别集合。那么由公式（7）计算条件概率。

$$p(y|\boldsymbol{S}) = \frac{e^{\sum_{i=1}^{n}Q_{i,y_i} + \sum_{i=1}^{n}A_{y_i,y_{i+1}}}}{\sum\limits_{y'\in Y^*} e^{\sum_{i=1}^{n}Q_{i,y_i} + \sum_{i=1}^{n}A_{y_i,y_{i+1}}}} \quad (7)$$

其中，$\boldsymbol{Q}$为概率矩阵，$Q_{i,\ yi}$表示标记序列中第i个位置为y_i的概率；$\boldsymbol{A}$是状态转移矩阵，$A_{yi,\ yi+1}$表示由标记y_i到标记y_{i+1}的概率；Y^*是对于输入序列S而言，所有可能的标记序列集合。初始$\boldsymbol{Q}$由通过CRF中的全连接层转化得来，$\boldsymbol{A}$则由CRF模型随机初始化得来。概率矩阵$\boldsymbol{Q}$和转移矩阵$\boldsymbol{A}$共同决定如何选择标记序列，得分最大的路径作为最终的结果。

2.4 训练阶段

公式（7）中决定最终标记序列的 $\boldsymbol{Q}$ 和 $\boldsymbol{A}$ 是可学习参数，可通过对模型的训练得到。训练阶段的loss损失函数选择条件概率的对数似然函数，如公式（8）所示。

$$L(\boldsymbol{S}, y) = -\log(p(y|\boldsymbol{S})) \tag{8}$$

在反向传播过程中根据损失函数来不断调整 $\boldsymbol{Q}$ 和 $\boldsymbol{A}$，最终生成约束标签先后顺序关系的转移矩阵 $\boldsymbol{A}$；本研究选择Adam优化算法学习模型参数。

当学习到 $\boldsymbol{Q}$ 与 $\boldsymbol{A}$ 矩阵后，在预测阶段，通过维特比算法求解条件概率 p（$y|\boldsymbol{S}$）最大时的标记序列 y*作为输入序列的标注，即公式（9）。

$$y^* = \underset{y \in Y^*}{\operatorname{argmax}}\ p(y|\boldsymbol{S}) \tag{9}$$

3 模型有效性验证与实验设计

为验证本方法的有效性，分别选择预训练BERT、NEZHA[23]以及BERT-wwm[24]作为模型表示层。这些模型均是具有12层Transformer[11]结构的基础模型。将输入序列通过上述预训练语言模型进行上下文学习得到模型对应的表示层集合，再添加本研究提出的递进式卷积网络模块至预训练语言模型和CRF解码层之间，得到增强的句子聚合表示层进行下一步命名实体识别，观察其评价指标。

3.1 数据获取与评价指标

在公开数据集上，选取人民日报的PeopleDaily命名实体识别数据集和微软亚洲研究院的MSRA命名实体识别数据集。PeopleDaily和MSRA数据集均将命名实体分为人名、地名、组织名三类。数据集被划分为训练集和测试集，相关信息如表1所示。

在农业领域，缺少公开标注的数据集[14]，于是在现有的农产品本体知识库中人工整理收集数据，构建农产品命名实体识别数据集AgriNER。本体中的概念是对实例的抽象，指向某一类性质相同或相近的实例。因此，本研究将本体中的所有概念、实例统一当作命名实体，信息如表2和表3所示，该任务识别出农产品命名实体，并将其分为五类，分别是农产品类别（Product Class，PC）、农产品实例（Product Instance，PI）、病虫害类别（Disease and Pest Class，DPC）、病虫害实例（Disease and Pest Instance，DPI）和行政区划（Region，RI）。

表1 公开命名实体识别数据集统计信息

Table 1 Public Named Entity Recognition（NER） dataset statistics

数据集	类型	训练集/个	训练集/个
PeopleDaily	句子数量	20,864	4346
	实体数量	33,992	7707
MSRA	句子数量	45,000	3442
	实体数量	7559	6192

表2 AgriNER数据集统计信息

Table 2 AgriNER dataset statistics

统计对象	训练集/个	验证集/个	测试集/个	总数/个
句子数量	5050	1682	1682	8414
实体数量	4351	1456	1449	7256
农产品类别	157	35	38	230
农产品实例	1140	387	373	1900
病虫害类别	103	33	22	158
病虫害实例	1099	406	372	1877
行政区划	1852	615	644	3111

表3 AgriNER 标注映射表

Table 3 AgriNER annotation mapping table

实体类型	实体首字符标注	实体非首字符标注
农产品类别	B-Product-Class	I-Product-Class
农产品实例	B-Product-Instance	I-Product-Instance
病虫害类别	B-DP-Class	I-DP-Class
病虫害实例	B-DP-Instance	I-DP-Instance
行政区划	B-Region-Instance	I-Region-Instance

实验采用BIO标注法对实验数据进行标注。其中，B表示实体的开始，即命名实体的首字符标注为B；I表示实体的其余部分，即命名实体中除首字符的其余字符标记为I；O表示非实体部分，即句子中的非实体字符均标记为O。

为了评价方法在命名实体识别任务数据集上的表现，选取命名实体识别任务常用的评价指标，包括精确率 P，召回率 R 和 F_1 值。F_1 值为 P 和 R 的调和平均数，评价模型的整体性能。

3.2 实验环境与参数设置

实验环境如表4所示，分别以BERT、NEZHA和BERT-wwm作为模型表示层，由于构建的农业语料库相对公开数据集规模较小，该模型先在公开数据集上进行实验测试，同时确定卷积核大小以及预训练模型等超参数。

通过多次实验确定模型超参数，如表5所示。此外，根据表示层选择不同，训练轮次epoch大小不同。其中除了当数据集为MSRA，并且表示层选用BERT时，epoch设置为3，其余实验epoch均为5。

表4 NER实验环境

Table 4 NER experimental environment

操作系统	Windows 10
CPU型号	Intel Xemon CPU E5-1630 v4 @3.70 GHz
GPU型号	Titan X
Python版本	3.7
Tensorflow版本	1.14
内存大小	64 GB

表5 NER实验模型超参数

Table 5 NER experimental model hyperparameters

参数	值
最大序列长度	128
批大小	32
学习率	0.00005
Dropout失活率	0.5
卷积核大小	5×5

4 实验结果分析与讨论

4.1 公开数据集实验结果分析

实验均分别重复3次，取平均值作为实验结果，以中和模型参数随机初始化的影响，结果如表6所示。其中选取的BERT-BiLSTM[7]可作为BERT与原先深度神经网络进行结合的基线模型，而Sesame[19]和JAM[18]模型是采用另外两种多层表示聚合模型。

从表6中可以发现，本研究提出的基于递进式卷积网络的命名实体识别方法在PeopleDaily和MSRA数据集上的F_1值均高于其他模型，并且其表现整体优于Sesame和JAM多层表示融合模型。相较于BERT模型，本方法在PeopleDaily数据集上的F_1提升0.51%，在MSRA数据集上F_1提升0.84%。实验结果表明，本方法能够一定程度上增强模型对自然语言的表示能力，提升模型在命名实体识别任务上的准确率。

表6 模型在公开数据集NER试验结果对比——以BERT为基础预训练模型

Table 6 Comparison of the NER experimental results on public datasets — based on BERT

模型	PeopleDaily			MSRA		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
BERT	93.81	94.12	93.97	94.48	93.78	94.12
Sesame	86.05	85.53	85.79	88.76	87.18	87.96
JAM	90.25	90.88	90.57	90.47	91.52	90.99
BERT-BiLSTM	93.77	94.36	94.07	94.16	87.18	94.55
本文模型	94.53	94.44	94.48	94.04	94.89	94.96

注：P为精确率，R为召回率，F_1值为P和R的调和平均数

表7和表8分别是基于NEZHA和BERT-wwm模型的实验结果。可以看出，本研究的递进式卷积网络融合方法在NEZHA和BERT-wwm上有显著效果。其中，基于NEZHA模型在PeopleDaily数据集上F_1提升0.19%，在MSRA数据集上F_1提升0.23%；基于BERT-wwm语言模型在PeopleDaily数据集上F_1提升0.24%，在MSRA数据集上F_1提升0.53%。从实验结果可以看出，本方法不仅对BERT模型有效，对于与BERT有相同结构的预训练语言模型有相同的作用。

4.1.1 对不同表示层的影响分析

从表6~8的实验结果可以发现，本方法在

表7 模型在公开数据集NER试验结果对比——以NEZHA为基础预训练模型

Table 7 Comparison of the NER experimental results on public datasets — based on NEZHA

模型	PeopleDaily			MSRA		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
NEZHA	95.11	94.80	94.96	95.33	95.23	95.28
Sesame	90.11	90.19	90.15	92.79	92.29	92.53
JAM	92.73	92.58	92.65	93.59	93.52	93.55
本文模型	94.92	95.37	95.15	95.21	95.82	95.51

表8 模型在公开数据集NER试验结果对比——以BERT-wwm为基础预训练模型

Table 8 Comparison of the NER experimental results on public datasets — based on BERT-wwm

模型	PeopleDaily			MSRA		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
BERT-wwm	93.56	94.19	93.88	93.80	94.29	94.04
Sesame	84.85	84.63	84.66	90.00	90.42	90.20
JAM	91.02	90.64	90.82	92.04	92.99	92.51
本文模型	94.01	94.23	94.12	94.55	94.59	94.57

BERT上的成绩提升效果明显优于在BERT-wwm和NEZHA模型上的成绩提升，如图4所示。这种差异性取决于语言模型自身特点：BERT-wwm和NEZHA模型是在BERT模型的基础上，采用全词掩码进行改进，并且NEZHA在注意力矩阵中添加了相对位置编码。从信息论的角度出发，BERT-wwm和NEZHA在BERT基础上的优化，增强了模型对信息的编码能力，从而降低了编码过程的不确定性，减少了表示聚合能增强的信息量，因此三者通过递进式卷积网络实现的表示聚合的效果提升有上述层次差异。

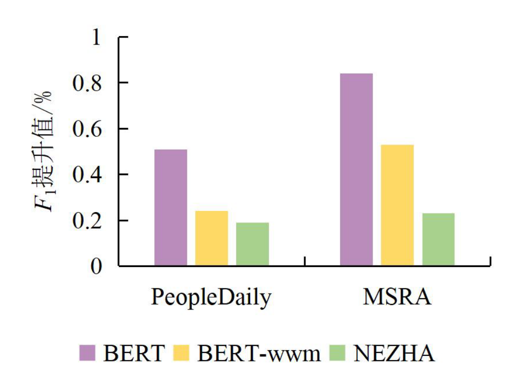


图4 公开数据集NER实验不同模型F_1值提升对比

Fig. 4 Comparison of F_1 values of different models in the NER experiment on the public datasets

4.1.2 不同融合方法对比

从实验结果中可以发现，Sesame模型和JAM模型的实验结果均差于原模型，其中JAM模型在融合解码器每层表示时只简单进行线性组合，更多的是借鉴门控网络调节选择流向下一层的信息量，需要自行学习权重；Sesame则用挤压融合和激励操作获取BERT各层的权重因子，按照权重对全层输出进行加权获得最终表示。上述两种BERT的多层表示融合方法的核心思想均是通过对BERT各层进行加权，进而融合加权后的各层分布式表示，将权重的学习也交付给模型本身进行。权重根据BERT各层在下游任务数据上表现能力，强化更能适应任务的中间层表示，弱化不重要的中间层。但强化或弱化某一中间层全部的分布式表示，在某种程度上还是忽略了一部分信息。而本研究提出的融合方法平均保留了全层的信息，相对限制了模型的学习自由度，强制模型重视每层信息，能够更好地聚合预训练语言模型的各层嵌入，更多地挖掘模型在大规模语料上学习到的语言特性。

4.1.3 卷积操作以及核尺寸影响

为了探索卷积层中卷积操作以及核尺寸对于模型的影响，实验以BERT作为表示层，进行了当卷积核尺寸为5×768的实验。其中，768表示BERT的隐层维度。本研究选取该卷积核尺寸的原因在于，在分类等自然语言处理任务中，这种尺寸的卷积核有较好的效果。实验结果如表9所示，表中参数量为在BERT的原参数量110 M的增加量，T_e表示单轮次训练时间，表格第一行为BERT模型的实验记录。

表9 实验卷积操作以及核尺寸影响

Table 9 Convolution operations and kernel size effects

卷积核尺寸	参数量	PeopleDaily		MSRA	
		F_1/%	T_e/min	F_1/%	T_e/min
/	/	93.97	15	94.12	28
5×5	300	94.48	18	94.96	30
5×768	46080	93.48	150	94.53	290

实验结果表明，小尺寸的卷积核更适用于本方法。首先，大尺寸的卷积核在F_1值上没有提升，说明没有实现有效的表示聚合。其次，由于卷积核尺寸的增大，大尺寸的卷积层参数量更大，因此需要更多的训练时间。所以，本研究提出的递进式卷积网络在命名实体识别任务上更适合采用小尺寸卷积核。与BERT模型相比，本模型未有显著的时间和空间需求增加，可见增加递进卷积模块后的模型空间复杂度以及时间复杂度在可承受范围。

4.2 农产品数据集实验

由4.1节实验及结果分析可以知道，以预训练

语言模型BERT为基线的模型性能提升最有效，且卷积核尺寸对模型F_1值有明显作用，因此，农产品命名实体识别以BERT为基线实验，分别设置不同尺寸卷积核进行实验以分析本方法的作用，实验结果如表10。

表10 AgriNER数据集实验结果

Table 10 AgriNER experimental results

模型	P/%	R/%	F_1/%
BERT	90.00	83.09	85.96
本文模型(3×3)	88.73	86.13	86.69
本文模型(5×5)	88.72	90.45	89.57
本文模型(7×7)	87.68	89.19	88.41

4.2.1 农产品命名实体识别实验结果分析

实验发现，本方法相比于传统方法能够显著提升模型R和F_1值。首先，当卷积核尺寸为5×5时，F_1值提升最大，且当卷积核尺寸为3×3时，F_1值最小，但是优于传统方法。R在公开数据集上也表现出同样情况。其次，可以发现本方法对P存在影响。当卷积核尺寸越大时，P值越小，但是降低幅度小于R的提升幅度。

4.2.2 不同类别命名实体上的表现分析

为了分析模型在不同类别命名实体上的表现，本研究统计了模型在AgriNER数据集上各个实体类别的评价指标，发现R和P情况与F_1值相似，故选取F_1值进行说明。F_1值统计结果如表11所示，传统方法BERT在各类实体上的指标值波动较大（53.73～98.69），且对行政区划实体F_1值最高，对农产品类别F_1值最低。其次，本方法当卷积核尺寸为5×5时，F_1值在各类实体上的指标波动最小（71.60～98.87），且对行政区划实体F_1值最高，对农产品类别实体F_1值最低。此外，对于所有模型，对农产品类别实体的F_1值都是最低的。造成这种情况的原因是农产品类别和农产品实例的实体相似度较高。比如“豌豆类”是一个农产品类别，包括的农产品实例有“豌豆”“鲜豌豆米”“荷兰豆”等。同样的，病虫害类别的F_1值也整体低于病虫害实例F_1值也是相同原因。

通过实验结果分析可看出，本方法不仅优于传统基于BERT的命名实体识别方法，也优于当前基于BERT模型的其余各种融合方法。其中，卷积核尺寸对模型影响较为明显，当卷积核尺寸取中间值5×5时，模型表现最佳，稳定性最强。因此可以得出结果，本方法能够有效地增强预训练语言模型对于命名实体的表示和识别能力，且卷积核尺寸不应当过大或过小。

表11 AgriNER数据集各类实体NER实验F_1值比较

Table 11 Comparison of F_1 values for NER experiments of various entities on the AgriNER datasets

模型	实体类别				
	PC	PI	DPC	DPI	RI
BERT	53.73	91.88	90.91	94.61	98.69
本文模型(3×3)	58.46	92.20	89.80	94.22	98.78
本文模型(5×5)	71.60	92.86	88.89	95.64	98.87
本文模型(7×7)	64.20	91.21	93.88	94.22	98.55

注：农产品类别(Product Class，PC)、农产品实例(Product Instance，PI)、病虫害类别(Disease and Pest Class，DPC)、病虫害实例(Disease and Pest Instance，DPI)和行政区划(Region，RI)

4.2.3 分析标记数据对实验评估性能影响

由精度的公式P=TP/（TP+FP）出发，分析实验结果的TP（真正例）和FP（假正例）。统计发现与BERT模型相比，本模型在TP上与BERT模型未有明显数量差异，三次实验中TP数量差距均在10之内，约占据平均TP数据量的0.7%，而在FP上比BERT模型统计量多于10，约占据平均FP数据量的25%，从而导致精度P在数值上低于BERT模型。对假正例进行人为观察，发现模型测得的假正例实体有部分是属于自然语言实体，但未出现在测试集标签中。由于在对本体农业知识库进行标注时对句子进行随机字符替换，使得数据集有一定噪声。类比非嵌套公开数据集MSRA也是如此。可能本模型过多地融合了浅层信息，使得实体识别灵敏度过高，未匹配上标签集给定的实体名称或长度。同时从侧面说明了标记质量与策略对模型评估的重要性。

5 结论

本研究针对现有基于预训练语言模型的命名实体识别方法忽视预训练语言模型内部各层本身蕴含语言不同层次的丰富信息，仅使用模型最后一层表示输出，对模型利用不充分的问题，提出递进式卷积网络，聚合预训练语言模型所有编码层输出的句

子表示。方法利用递进式卷积操作提取语言模型蕴含的不同维度的语言信息，同时注重模型浅层和深层信息，使模型在拥有深度语义信息的同时，融合对农业实体名称有利的粗粒度信息，增强模型对句子实体的表示能力。相较于BERT模型，本文方法在PeopleDaily数据集上的F_1提升0.51%，在MSRA数据集上F_1提升0.84%，在AgriNER数据集上F_1提升3.61%。实验结果表明，本方法不仅在公开数据集上能够有效提升模型准确率，同时在农业领域的实际应用中能够更有针对性地定位实体位置，一定程度上解决专业名词的识别边界模糊，识别率不高的问题。同时验证了小尺寸卷积核的有效性，结果表明利用递进式网络去增强表示从而提升命名实体识别准确率是有效的。

但在非嵌套型实体名称的长度上，模型融合了过多浅层信息，在上下文语义信息的获取上还有提升空间。由于语义信息也可作为实体名称长度截断点选取的重要参考信息，在以后的研究中可以进一步思考如何在保持浅层信息融合的同时对深层信息进行增强。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献：

[1] QIU X P, SUN T X, XU Y G, et al. Pre-trained models for natural language processing: A survey[J]. Science China technological sciences, 2020, 63(10): 1872-1897.

[2] SEVASTJANOVA R, KALOULI A, BECK C, et al. Explaining contextualization in language models using visual analytics[C]// 2021 59th Association for Computational Linguistics (ACL). Stroudsburg, PA, USA: Association for Computational Linguistics, 2021: 464-476.

[3] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4171-4186.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// 2017 31st Annual Conference on Neural Information Processing Systems (NIPS). La Jolla, California, USA: Neural Information Processing Systems, 2017: 6000-6100.

[5] 杨飘, 董文永. 基于BERT嵌入的中文命名实体识别方法[J]. 计算机工程, 2020, 46(4): 40-45, 52.
YANG P, DONG W Y. Chinese named entity recognition method based on BERT embedding[J]. Computer engineering, 2020, 46(4): 40-45, 52.

[6] GAN Y, YANG R S, ZHANG C F, et al. Chinese named entity recognition based on BERT-transformer-BiLSTM-CRF model[C]// 2021 7th International Symposium on System and Software Reliability (ISSSR). Piscataway, NJ, USA: IEEE, 2021: 109-118.

[7] GAO W C, ZHENG X H, ZHAO S S. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF[J]. Journal of physics. Conference series. 2021, 1848(1): ID 012083.

[8] CHANG Y, KONG L, JIA K J, et al. Chinese named entity recognition method based on BERT[C]// 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA). Piscataway, NJ, USA: IEEE, 2021: 294-299.

[9] LI X, YAN H, QIU X, et al. FLAT: Chinese NER Using Flat-Lattice Transformer; proceedings of the ACL, F, 2020[C]// 2020 58th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6836-6842.

[10] 琚生根, 李天宁, 孙界平. 基于关联记忆网络的中文细粒度命名实体识别[J]. 软件学报, 2021, 32(8): 2545-2556.
JU S G, LI T N, SUN J P. Chinese fine-grained name entity recognition based on associated memory networks[J]. Journal of software, 2021, 32(8): 2545-2556.

[11] WANG X Y, JIANG Y, BACH N, et al. Improving named entity recognition by external context retrieving and cooperative learning[J/OL]. arXiv: 2105.03654, 2021.

[12] NIE Y Y, TIAN Y H, SONG Y, et al. Improving named entity recognition with attentive ensemble of syntactic information[C]// Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 4231-4245.

[13] 李林, 周晗, 郭旭超, 等. 基于多源信息融合的中文农作物病虫害命名实体识别[J]. 农业机械学报, 2021, 52(12): 253-263.
LI L, ZHOU H, GUO X C, et al. Named entity recognition of diseases and insect pests based on multi source information fusion[J]. Transactions of the Chinese society for agricultural machinery, 2021, 52(12): 253-263.

[14] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于注意力机制的农业文本命名实体识别[J]. 农业机械学报, 2021, 52(1): 185-192.
ZHAO P F, ZHAO C J, WU H R, et al. Named entity recognition of Chinese agricultural text based on attention mechanism[J]. Transactions of the Chinese society for agricultural machinery, 2021, 52(1): 185-192.

[15] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language? [C]// 2019 57th Annual Meeting of the Association for Computational Linguistics (ACL). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3651-3657.

[16] ROGERS A, KOVALEVA O, RUMSHISKY A. A primer in BERTology: What we know about how BERT

works[J]. Transactions of the association for computational linguistics, 2020, 8: 842-866.
[17] JIE Z M, LU W. Dependency-guided LSTM-CRF for named entity recognition[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4231-4245.
[18] ZHANG Z B, WU S, JIANG D W, et al. BERT-JAM: Maximizing the utilization of BERT for neural machine translation[J]. Neurocomputing, 2021, 460: 84-94.
[19] SU T C, CHENG H C. SesameBERT: Attention for anywhere[C]// 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Piscataway, NJ, USA: IEEE, 2020: 363-369.
[20] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 42(8): 2011-2023.
[21] JIANG Z, YU W, ZHOU D, et al. ConvBERT: Improving BERT with Span-based Dynamic Convolution[J/OL]. arXiv:2008.02496 [cs.CL], 2020.
[22] SUN C J, GUAN Y, WANG X L, et al. Rich features based Conditional Random Fields for biological named entities recognition[J]. Computers in biology and medicine, 2007, 37(9): 1327-1333.
[23] WEI J, REN X, LI X, et al. NEZHA: Neural contextualized representation for Chinese language understanding[J/OL]. arXiv:1909.00204v3 [cs.CL], 2009.
[24] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM transactions on audio, speech and language processing, 2021, 29: 3504-3514.

Progressive Convolutional Net Based Method for Agricultural Named Entity Recognition

JI Jie[1,2], JIN Zhou[1], WANG Rujing[1,2*], LIU Haiyan[1,2], LI Zhiyuan[1,2]

(1. *Institute of Intelligent Machinery, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei* 230031, *China*; 2. *University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: Pre-training refers to the process of training deep neural network parameters on a large corpus before a specific task model performs a particular task. This approach enables downstream tasks to fine-tune the pre-trained model parameters based on a small amount of labeled data, eliminating the need to train a new model from scratch. Currently, research on named entity recognition (NER) using pre-trained language model (PLM) only uses the last layer of the PLM to express output when facing challenges such as complex entity naming methods and fuzzy entity boundaries in the agricultural field. This approach ignores the rich information contained in the internal layers of the model themselves. To address these issues, a named entity recognition method based on progressive convolutional networks has been proposed. This method stores natural sentences and outputs representations of each layer obtained through PLM. The intermediate outputs of the pre-trained model are sequentially convolved to extract shallow feature information that may have been overlooked previously. Using the progressive convolutional network module proposed in this research, the adjacent two-layer representations are convolved from the first layer, and the fusion result continues to be convolved with the next layer, resulting in enhanced sentence embedding that includes the entire information dimension of the model layer. The method does not require the introduction of external information, which makes the sentence representation contain richer information. Research has shown that the sentence embedding output of the model layer near the input contains more fine-grained information, such as phrases and phrases, which can assist with NER problems in the agricultural field. Fully utilizing the computational power already used, the results obtained can enhance the representation embedding of sentences. Finally, the conditional random field (CRF) model was used to generate the global optimal sequence. On a constructed agricultural dataset containing four types of agricultural entities, the proposed method's comprehensive indicator F_1 value increased by 3.61% points compared to the basic BERT (Bidirectional Encoder Representation from Transformers) model. On the open dataset MSRA, the F_1 value also increased to 94.96%, indicating that the progressive convolutional network can enhance the model's ability to represent natural language and has advantages in NER tasks.

Key words: agriculture named entity recognition (NER); pre-trained language model (PLM); convolutional net; representation aggregation; deep learning